MASSACHUSETTS INSTITUTE OF TECHNOLOGY

ARTIFICIAL INTELLIGENCE LABORATORY

Memo No. 327 December 1974

A NOTE ON THE COMPUTATION OF BINOCULAR DISPARITY IN A SYMBOLIC, LOW-LEVEL VISUAL PROCESSOR

by

David Marr

ABSTRACT

The goals of the computation that extracts disparity from pairs of pictures of a scene are defined, and the constraints imposed upon that computation by the three-dimensional structure of the world are determined. Expressing the computation as a grey-level correlation is shown to be inadequate. A precise expression of the goals of the computation is possible in a low-level symbolic visual processor: the constraints translate in this environment to pre-requisites on the binding of disparity values to low-level symbols. The outline of a method based on this is given.

Work reported herein was conducted at the Artificial Intelligence Laboratory, a Massachusetts Institute of Technology research program supported in part by the Advanced Research Projects Agency of the Department of Defense and monitored by the Office of Naval Research under Contract number N00014-70-A-0362-0005.

Introduction

Commercial pressures have led to considerable interest in the automatic extraction of disparity information from pairs of pictures of a scene. Since 1968, there has been available a machine, the Wild-Raytheon B8 stereomat automated plotter, which can draw a contour map from two aerial photographs. The machine correlates intensity measurements obtained over local scans made on the two images: the scan paths are the machine's current approximation to the contour lines (i.e. lines of constant disparity), and the failures of correspondence between the scans on the two images are used to improve the approximation. Adequate accuracy, if achieved at all, is reached within about six iterations.

Machines that seek to assign disparity values to an image by performing correlations between intensity arrays are subject to troublesome problems due to local minima: Mori, Kidode & Asada (1973) describe the problems, and have recently implemented some ways of avoiding them. Their principal cures are (i) to correlate the two images using local averages taken over regions that are initially relatively large, and which are subsequently reduced in size as the solution is approached; and (ii) to avoid local minima traps by introducing a small amount of gaussian noise into the images. These techniques reduce considerably the incidence of false assignments, but they fail to remove them altogether.

There has been less progress in the study of parallel algorithms

for making use of disparity information, despite considerable recent interest in the processing of disparity information by the visual system, (Barlow, Blakemore & Pettigrew (1967), Julesz (1971)): Julesz (1971 p204), and Sperling (1970) have both suggested possible schemes. Julesz's model is informal in nature, being more phenomenological than computational: it is very interesting because despite its great simplicity, it displays an astonishing number of the properties that are exhibited by the human disparity processing system. Sperling's model is more formal, but it is difficult to tell how well it would work. This is a problem with all complex parallel methods; they are very expensive to simulate, and it is extremely difficult to derive analytically, from a system with complex non-linear components, quantities that could be measured experimentally. About the best one can hope to do at present is to state criteria that distinguish one family of methods from another, and ask whether those criteria are satisfied by the particular method that we use.

This note enquires about the exact nature of the disparity computation. It is in some sense a correlation, and because that is common knowledge and fairly precise, a deeper characterisation has not been sought. But in order to formulate a method of carrying it out, one needs to be very precise about the goals of the computation, and about the constraints imposed upon it by the structure of the three-dimensional world. Unless one uses a method that is based on all and only the correct assumptions, there will be situations in which it will fail unnecessarily.

Measuring disparity

If a scene is photographed from two slightly different positions, the relative positions of the objects in the scene will differ slightly on the two images. The discrepancies of interest arise from the different distances of the objects from the viewing position, and measurements of the discrepancies contain useful information about the relative distances of the objects. The term binocular disparity refers to the difference in the angle from each eye to a point in the scene, measured relative to some suitably chosen angle of convergence. The central difficulty in defining what is meant by the process of extracting binocular disparity from an image is that disparity has to refer to a physical entity - a point on a visible surface - yet it appears that we compute it at a level far below that at which the world is described in terms of surfaces and objects (Julesz 1971). It was probably this fact that made so surprising Julesz's conclusion that disparity assignment is a low-level computation.

In order to compute disparity correctly, the following steps must be carried out: firstly, a particular location on a surface in the scene must be located in one image; secondly, the identical location must be identified in the other image; thirdly, the relative positions of the two images of that location must be measured. The most interesting and most troublesome part of the process concerns the selection of a location on the viewed surface, and the identification of its two images. The difficulty is that the choice of a point on the surface must be made from the images: if it could be chosen in some absolute way - by lighting it

up at that point for example - the problem would be simple.

We are now in a position to understand why the disparity computation is not the same thing as a grey-level correlation. The reason is that grey-level measurements correspond to properties of the image, rather than to properties of the objects being viewed. An (x, y) co-ordinate pair on an image is an artefact of the transducer, since it does not define a point on a physical surface in a way that allows it to be identified on the other image. The most glaring example of the failure is the case where an image point corresponds to two surface points, the nearer of which is transparent or translucent: a goldfish in a pond is one such case, where the water surface and the goldfish are simultaneously visible. Other examples are provided by figures 5.7.1 and 6.3.2 of Julesz (1971). But the argument applies equally well to the case of a single visible surface, and its consequence is that grey-level matching methods are incorrect. On simple images, the method will succeed, because it is close enough to the right idea: but as Mori et al. (1973) have found, it will not succeed on complex images because it is based on incorrect premises. Their technique of introducing local smearing may be viewed as a way of beginning to identify a point in the image with a point on the physical surface (by adding additional constraints on what is matched): in so far as it does so, the method will become more reliable, but it is probably better to attack the underlying issue directly.

The use of low-level symbols

In order to formulate the disparity computation in a usable way, we therefore need to be able to identify surface points from the two images, and match them up. It is clearly fruitless to try to label points of a smooth featureless surface, but if that surface contains a scratch, boundary, or other identifying physical mark which produces a local and fairly sharp change in reflectance, that change in reflectance may be used to define the surface point. Provided that the change in reflectance has been identified and described separately in the two images, the resulting descriptions will correspond to an underlying physical reality. The computation of such a low-level description from one image has been dealt with at length elsewhere (Marr 1974a & 1974b), and is called the low-level symbolic representation of an image. Hence we see that provided stereo matching takes place between two low-level symbolic descriptions, it is a well-founded operation.

Finally, we need to ask whether any reasonably complex measurement could be used - or is there something special about a low-level symbolic description? Simple cell-like measurements are for example nearly suitable, because they are sometimes quite near to low-level assertions: but when assigning disparity values to simple cells, one meets all the usual problems associated with measurements - a whole set of simple cells, at neighbouring positions and orientations, corresponds to the underlying scratch or whatever in the image, and it is that complex which needs to be matched against the corresponding complex derived from the other image. If the important matching step is carried

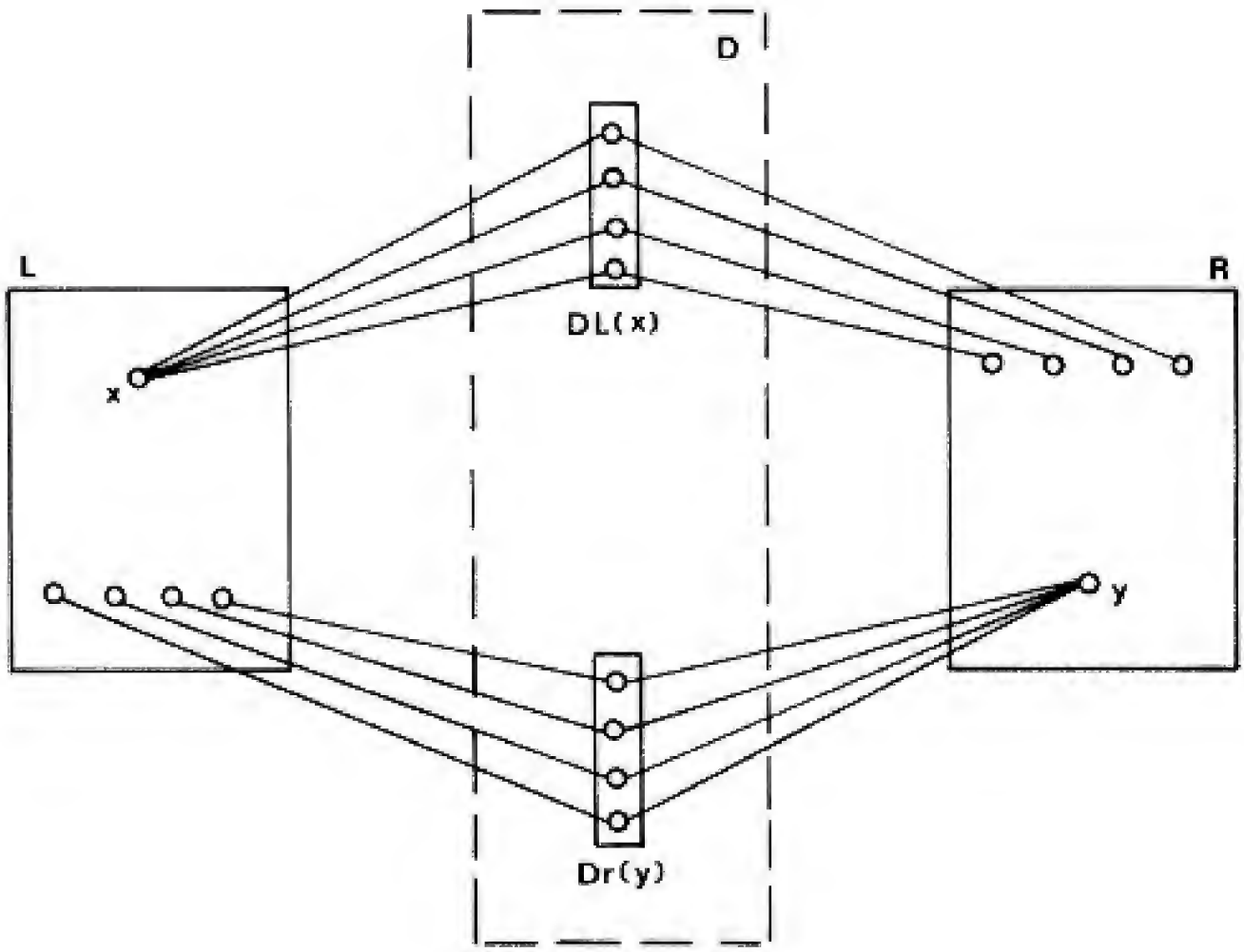

FIGURE 1

Figure 1. L denotes the collection of left-image symbols, and R the collection of right-image symbols. These are connected through the set D of disparity value symbols. The sets DL and Dr referred to in the text are shown in the figure.

out on each individual simple cell measurement, the computation becomes very uneconomical. Hence one may expect that when disparity is actually assigned, the process operates on a very low-level symbolic description. This method will fail only when the low-level decriptions obtained from the two images are very different: but this is comparatively rare, and one seems to notice it when it happens. In any case, this circumstance cannot be dealt with at the same very low level.

The problem has now been reduced to the comparason of two low-level symbolic descriptions, and the assignment of disparity values to pairs of symbols, drawn appropriately from each image. We turn now to examine briefly the rules and constraints to which this process is subject.

The "use once" condition

We have seen that an element of a low-level symbolic description of an image corresponds to a physically identifiable entity in a way that an image co-ordinate does not, and in which measurements made on an image only approximate. This allows one to state the first condition that controls the matching of two low-level symbolic descriptions. It is that each low-level symbol should be assigned exactly one disparity value, which in turn implies that it should be associated with at most one symbol computed from the other image. This is called the "use-once" condition, and it is non-linear.

The use-once condition may be implemented in the following way. Let L be the set of all left-image low-level symbols, and let R be the set of all right-image low-level symbols. To each element x in L, there

correspond several elements in R, one for each of the possible disparity values; and for each element y in R, there is a corresponding set of elements in L. This situation is illustrated in figure 1. The matching of one element from L with one from R corresponds to the assignment of a single disparity value to both elements, so let us introduce a third set D that consists of collections of elements representing all of the possible disparity values that may be bound to each low-level symbol. In principle, one needs one such collection for each low-level symbol, though members of L and R share elements in D in the appropriate way (see below). The use-once condition translates into the constraint that each element of L and of R may be bound to at most one element of D.

In practise, the set D will be very large unless steps are taken to economise on the number of units that are necessary to represent the disparity values; so let us consider how disparity representing units may be shared between several elements of L (say). D has to be large enough so that (a) each low-level symbol can find an unused collection in D that can be used for representing its disparity; and (b) the correspondence between L and R through D is well-defined. To accomplish this, the collection of left- (or right-) image symbols that share a given disparity-representing unit should have the property that it is very rare for two to be provoked simultaneously by the image.

When one considers how to construct a parallel network that implements the use-once condition, it is apparent that at least three variables must be accomodated: ascension and declination in the visual field, and disparity. A satisfactory arrangement is shown in figure 2: In

Figure 2. One possible geometrical arrangement of the set D of symbols for disparity values is to arrange D in stripes of constant disparity. The figure shows how this may be done. The stripes running nearly horizontally contain units with constant disparity di. The x-coordinate in the image is represented by the x-coordinate in the figure, and the y-coordinate in the image corresponds roughly to the y-coordinate in the figure. The y-axis is magnified in the figure because a given y coordinate in the image needs to be represented at each disparity. The point of the figure is to show how low-level left- and right-image symbols share elements of D. The sets Dl and Dr, that are described in the text and shown in figure 1, are arranged to cross one another, and divide D up in two different ways. The connexions that implement the use-once condition run along these two directions. The connexions that implement the suggestion interaction run more nearly parallel to the stripes, and they join stripes corresponding to the same disparity (di) in neighbouring positions in the visual field. The connexions that measure local goodness-of-fit would have an undirected distribution.

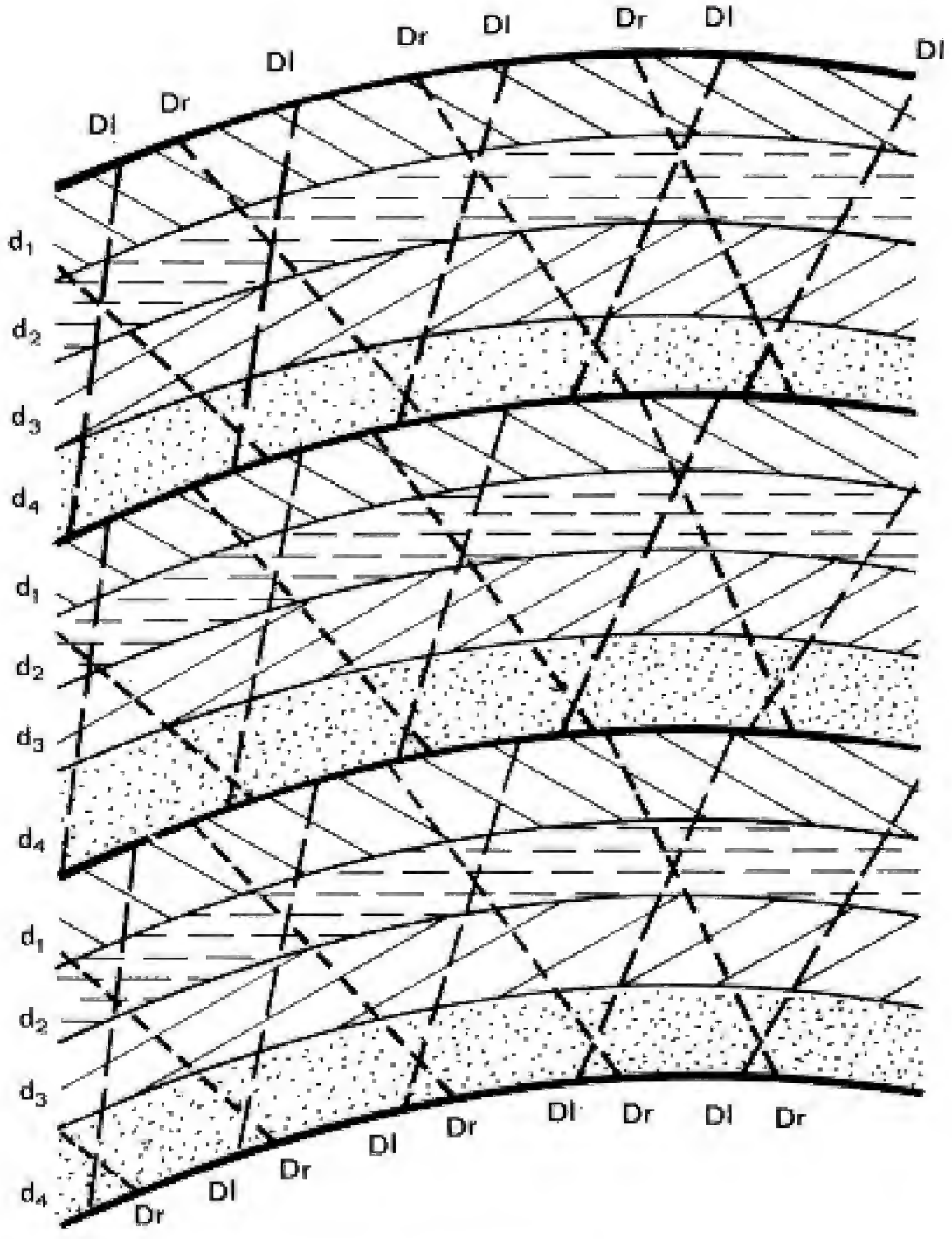
Dl
Dr
Dl
Dr
Dl
Dr
Dl
Dl
d1
d2
d3
d4
d1
d2
d3
d4
d1
d2
d3
d4
Dr
Dl
Dr
Dl
Dr
Dl
Dr
Dl

FIGURE 2

this, the units representing disparity values are arranged in stripes of constant disparity. The collections in D that represent disparity values for left-image symbols (L) lie along the diagonal lines marked Dl; and those for right-image symbols lie along the opposing set of lines Dr. Thus D is divided up in two ways into disparity representing units, which are simultaneously shared in an appropriate fashion by L and R. The connexions that implement the use-once condition are clearly marked: they run along Dl and Dr, joining places that contain representations of all possible disparity values that could be bound to a given left- or right-generated symbol. (In a neural implementation of this scheme, such connexions would be inhibitory.)

There are interesting differences between the implications for neurophysiology of these ideas, and of the model of Julesz. Firstly, the important part of the computation involves constraints on the disparity values that may be bound to low-level symbols. The magnets in Julesz's model seem however to correspond to rather unspecific local disparity values, and we saw earlier that the disparity computation cannot be accurately expressed in these terms. The second point rests on the way in which disparity-representing sets in D are assigned to L and to R. The most economical way of forming the collections of low-level symbols, that are to use the same disparity-representing units, is probably to group together all symbols that describe a small region and a small orientation range. Only very rarely will two such symbols be used simultaneously. If some scheme of this sort were being used, it would account for the existence of cells that behave sensitively to disparity, but are

relatively tolerant to position and orientation (Hubel & Wiesel 1970). Furthermore, it would be within these units that the main disparity computation is being carried out, and between which the governing connexions should be made. One would not expect to find other cells, expressing a free-floating disparity value in a "region" of the image, because the essence of the computation requires that it be carried out on bindings to low-level symbols. The model of Julesz would, I think, lead one to expect such cells.

Disparity is continuous almost everywhere

The use-once condition must be satisfied by the final assignment of disparity values to the low-level symbolic descriptions, but it is not much help in finding it. When applied to a random dot stereogram, it will ensure that the description of each dot, or group of dots, in one image is mated with not more than one similar description computed from the other image; and a solution that satisfies this and leaves very few dots out will probably be correct (see the next condition). But there is another useful property of the real world that can be introduced with advantage to speed the analysis. It is that except at object boundaries, disparity is a function that varies smoothly over the image. Fine texture, which is the best source of disparity information about a surface, will be represented at the lowest level by assertions about very small features; and except at object boundaries, the disparity values that become bound to neighbouring symbols will be about the same. This fact allows the existence of an interaction that "proposes" the disparity

used at one point as a strong candidate for the value at neighbouring points: it corresponds in Julesz's (1971) model to the lateral coupling provided by the small springs that join adjacent magnets. The implementation of this constraint in Sperling's (1970) model is obscure.

The implementation of a suggestion is one of those questions that it is not profitable to pursue in detail, because of the difficulty in testing, either physiologically or computationally, the conclusions to which one may be led. I shall therefore make only three points about it. The first is that, in principle, one would like a suggestion to influence the route to a solution, without disturbing the values in the solution once they are found. This implies a time-dependence in the interaction. Secondly, in order that the solution may be stable, one would probably also need to add a small DC component. The third point, and one which may actually be useful, concerns the geometry of the suggestion interactions, and this is shown in figure 2. There are connexions between all disparity units that represent similar disparity values, and that refer to symbols in nearby portions of the visual field. In a neural implementation, they would be excitatory.

Goodness-of-fit

The final important aspect of disparity measurement is the question of how satisfactory a solution is. Julesz emphasised the need for such a measure, and in his model, it corresponds to the total potential energy in the two, superimposed assemblies of magnets. Sperling (1970) also used a potential energy measure in his formulation. Julesz

showed that in an ambiguous stereogram, we perceive the better-correlated solution even if both have quite high correlations. This is good evidence that the matching process is parallel rather than serial, and that the goodness-of-fit measure is computed on a local basis.

In an implementation of the kind that we are discussing, the goodness-of-fit of a solution may be measured by the proportion of left- and right-image symbols that become bound to disparity values. In a perfect solution, the proportion will be 1.0; and an inappropriate disparity binding will have the effect of depressing the proportion of correct bindings in its neighbourhood. The local goodness-of-fit function would affect the confidence with which disparity assignments are made locally - i.e. the strength with which they are asserted - which in turn would affect the potency with which they are suggested to nearby regions. The goodness-of-fit function would therefore be implemented by a unit that depressed the local disparity-representing units by an amount that depended upon the proportion of image symbols in the vicinity that have been assigned disparity values. I do not see how to test for the presence of such units, except by trial-and-error.

Summary of the disparity interactions

The interactions described above are now drawn together, and the conditions that are necessary to an implementation of this kind are made explicit.

(1) The disparity assignment is made as a result of a matching operation performed on two low-level symbolic descriptions, computed independently

from the left and right images.

(2) The matching is implemented by applying conditions to the process of binding symbolic disparity descriptors to the low-level symbols. These constraints are the use-once condition, the suggestion interaction, and maximising the goodness of fit.

(3) The use-once condition requires interactions whose geometry appears in figure 2. These interactions inhibit the confidence of those assignments that they connect.

(4) The suggestion interactions have the geometry shown also in figure 2. They connect disparity descriptors that represent similar disparity values, and that are capable of being bound to low-level symbols referring to neighbouring positions in the visual field. Such interactions would probably have a time-dependent component as well as a DC component.

(5) Maximising the goodness of fit of a solution would have to be implemented by a local goodness-of-fit function, that measures the proportion of low-level symbols that have successfully been bound to a disparity value, and which affects the confidence level of the bindings in that local region.

Discussion

I shall not attempt in this note to derive any properties of the above system. I have been unable to make much progress with an analytical approach to the problem, and the amount of time required for a computational study is very large. The approach set out here does however

illustrate (a) how the disparity computation may be well-founded; (b) the importance of low-level symbols in the formulation; and (c) how the important constraints may at least in a general way be represented by connexions with a straightforward geometry. This kind of geometrical arrangement is one that it is becoming possible to detect.

The second large issue concerns the way in which disparity information may be used. It is one thing to assign disparity values to low-level symbols, and quite another to divide up an image into regions on the basis of disparity information alone, and compute a description of the spatial extent of each region. For example, in any stereogram, the orientation information associated with the small squares or groups of squares that are actually matched bears no relation to the orientation of the edge at which disparity changes. One way of computing the higher, induced edges would of course be to treat disparity like intensity, and subject its values to a process like that used to obtain a low-level symbolic description from an intensity array (Marr 1974b). This method seems somewhat clumsy, however, because disparity is not the only kind of information (excluding intensity) from which directions and boundaries may be computed locally. Texture changes are another example, and so are more abstract outlines, like the envelope of a sparse tree in winter, or the boundary defined by a row of small, separated bushes across a garden. One would like to know whether all of these problems may be dealt with by a single method that can describe configurations of "places" in an image - these places being identified by a rather simple kind of local measurement made on the relevant type of information. There seems to be

clear evidence (and a definite computational need) for such a mechanism - one of its main functions being to set up an orientation in the image at a point, to describe configurations of places relative to that orientation, and to influence the direction relative to which local shapes in an image are described. It is however far from clear whether one such mechanism would suffice to service all of the demands of this kind, or whether the slightly differing computational requirements force the existence of a number of separate, but similar mechanisms. The question will be raised elsuhere (Marr 1975).

Acknowledgement: Work reported herein was conducted at the Artificial Intelligence Laboratory, a Massachusetts Institute of Technology research program supported in part by the Advanced Research Projects Agency of the Department of Defense and monitored by the Office of Naval Research under Contract number N00014-70-A-0362-0005.

References

Barlow, H.B., Blakemore, C. & Pettigrew, J.D. (1967). The neural mechanism of binocular depth discrimination. J. Physiol. (Lond.), 193, 327-342.

Hubel, D.H. & Wiesel, T.N. (1970). Stereoscopic vision in macaque monkey. Nature, 225, 41-42.

Julesz, B. (1971). Foundations of Cyclopean Perception. Chicago: The University of Chicago Press.

Marr, D. (1975) Configurations, regions, and simple texture vision (in preparation).

Marr, D. (1974a) The purpose of low-level vision. MIT Artificial Intelligence Laboratory Memo 324.

Marr, D. (1974b) The low-level symbolic representation of intensity changes in an image. MIT Artifical Intelligence Laboratory Memo 325.

Mori, K., Kidode, M. & Asada, H. (1973). An iterative prediction and correction method for automatic stereocomparation. Computer graphics and image processing, 2, 393-401.

Sperling, G. (1970). Binocular fusion: a physical and neural theory. J. Am. Psychol., 83, 461-534.